Applying the summation model in audiovisual speech perception

Kaisa Tiippana, Ilmari Kurki, Tarja Peromaa

Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Finland

kaisa.tiippana@helsinki.fi, ilmari.kurki@helsinki.fi, tarja.peromaa@helsinki.fi

Abstract

Discrimination thresholds for consonants [k] and [p] were measured with the method of constant stimuli for auditory, visual and audiovisual speech stimuli. Summation in audiovisual thresholds was assessed using Minkowski metric. When Minkowski metric exponent *k* is 1, summation is linear. When *k*=2, summation is quadratic. As *k* increases, summation decreases. We found that *k* was 1.7 across 16 participants, indicating strong summation. There was some individual variation, *k* values being generally between 1 and 2. These findings confirm that multisensory enhancement is substantial in audiovisual speech perception. They also suggest that the amount of summation is not the same in all individuals.

Index Terms: audiovisual speech, discrimination threshold, summation model, multisensory integration

1. Introduction

It is well-known that seeing a talker's face improves speech perception [1]. However, how strong is this audiovisual enhancement? Usually it is determined by measuring the recognition accuracy of auditory and audiovisual speech at various levels of acoustic noise. The difference between these conditions is taken to reflect the enhancement, which is largest at intermediate levels of noise [2]. This measure quantifies the amount of visual influence on speech perception. However, it does not allow to unravel how the enhancement arises.

Quantitative models have explanatory power, they make predictions, and they can be tested experimentally. The enhancement in multisensory perception can be evaluated using the summation model, also called Minkowski metric:

$$s_{AV} = \left(s_A^k + s_V^k\right)^{\frac{1}{k}} \qquad (1)$$

where *s* refers to the intensity relative to the threshold (normalized threshold), *A*=auditory, *V*=visual, *AV*=audiovisual, and *k*=model parameter, which depends on the strength of summation, so that its value is inversely related to summation strength, i.e. the amount of enhancement. The model was originally developed to model visual summation [3], but it has also been used in multisensory contexts [4-8].

To apply the model, unisensory and audiovisual thresholds are measured, and the best-fitting value of parameter *k* is estimated. Assuming that the response of the mechanism is the product of the sensitivity of the mechanism and the stimulus intensity (i.e. linear transduction [3]), normalized i.e. relative intensity corresponds to the response of the underlying mechanism. The thresholds can be visualized by the summation square, where the abscissa plots the visual threshold, the ordinate plots the auditory threshold, and audiovisual thresholds are plotted in these coordinates (Fig. 1).

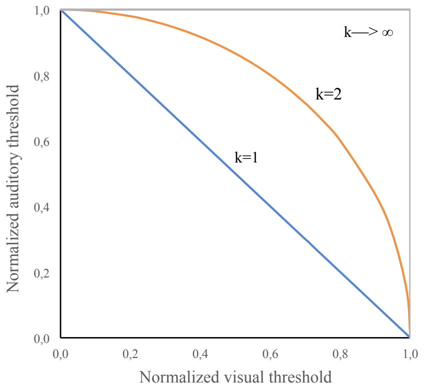

Figure 1: *Summation square. The plot shows the normalized audiovisual thresholds, i.e. auditory and visual components relative to the unisensory thresholds. Summation is linear when k is 1, and quadratic when k is 2. As k increases, summation decreases.*

In addition to quantifying the strength of audiovisual enhancement, the summation model offers explanations for it. The summation is the strongest when *k*=1. This means linear summation of unisensory thresholds. It suggests that auditory and visual signals are merged into the summation mechanism, and that the internal noise that limits performance arises after this summation. When *k*=2, summation is quadratic. This suggests that the signals have independent sources of noise, and are summed by combining the two sensory estimates, which are weighted according to their reliability. Higher values of *k* mean weaker summation.

In a multisensory context, the summation model has been used in studies of audiovisual motion perception. Linear summation has been found for directionally coherent, synchronous motion stimuli [4-5], but not always [6]. For the joint processing of spoken and written words, strong summation has also been found, with a *k* value of about 1.3 [7].

Audiovisual speech perception has previously been investigated using the Minkowski metric in a study by Arnold, Tear, Schindel and Roseboom [8]. They measured the discriminability of [b] and [g] in sentence context ("My name is Barry/Gary"). To be able to measure the discriminability index *d'*, they added white noise to the auditory stimuli and static black pixel masks to the visual stimuli. They compared the measured audiovisual *d'* to the predictions made by linear and quadratic summation of unisensory auditory and visual *d'*. In their six participants, naturally synchronous audiovisual speech discrimination performance was best accounted for by linear summation.

In the current study, thresholds were measured for discriminating two consonants embedded in meaningless syllables [ka] and [pa], using the method of constant stimuli. In this design, psychometric funtions (PFs) were obtained for all stimulus conditions, and the thresholds were defined at the 79 % correct level. Auditory thresholds were measured by varying sound intensity. Visual thresholds were measured by varying contrast. Audiovisual thresholds were measured at three intensity ratios of the unisensory components. The summation model was fitted to the AV thresholds normalized by the unisensory thresholds. The model fits were based on three AV thresholds instead of only one, as in [8], making them more robust. Since previous research has demonstrated strong summation for audiovisual speech and other multisensory stimuli, we expected to find linear or quadratic summation.

By using the summation model, it is possible to determine the strength of audiovisual enhancement, and to describe the mechanism producing the enhancement, as described above. Furthermore, the model can be fitted to each participant's data, providing individual estimates of summation. It is sometimes assumed that a single model can account for summation for all participants and situations [9]. The Minkowski metric allows to test whether this is the case by obtaining a quantified metric (*k*) of summation strength.

2. Methods

2.1. Participants

There were 16 participants (mean age 22.6 years, range 20-28; 4 male). They had normal or corrected-to-normal vision, no reported hearing or neurological disorders, and Finnish was their native language. This research was approved by the University of Helsinki Ethical Review Board in the Humanities and Social and Behavioural Sciences.

2.2. Equipment

The experiments were conducted using a standard computer run in Matlab environment, using Psychophysics Toolbox 3.10 extensions [10-11]. The visual stimuli were displayed on a 19" CRT screen (Sony G420). The screen resolution was 1024 x 768 px, size of active area 34.0 x 24.5 cm, and refresh rate 100 Hz. The maximum luminance of the was adjusted to 100 cd/m^2 and the gamma to 2.4. The sound card was Creative Sound Blaster X-Fi Titanium Fatal1ty Professional series, with a sample rate of 48000 Hz. ASIO driver was used for high precision auditory timing. The auditory stimuli were delivered via Beyerdynamic DT 770 Pro headphones.

2.3. Stimuli

The stimuli were five different iterations of five Finnish speakers (3 females) uttering the syllable [pa] or [ka]. The auditory and visual stimulus components were extracted from the same audiovisual video clips. The stimuli were presented in white noise.

The auditory stimuli had a duration of 166 ms on average. White noise at 31 dB SPL was superimposed onto the acoustic speech stimulus. The duration of the noise mask was 1 s.

The visual stimuli had a frame rate of 25 frames/s. They were converted into gray scale and windowed so that only the mouth region was visible (3.0 x 2.0°). A dynamic white noise mask (6.5 x 6.5°) was superimposed onto the visual stimulus. The root-mean-square contrast of the noise was 0.05. The visual stimulus presentation lasted 1 s.

Auditory and visual thresholds were measured by varying the intensity and contrast level of the stimuli, respectively. Natural synchrony was preserved in audiovisual conditions.

2.4. Procedure

In a 2-interval forced-choice paradigm, the task of the participant was to discriminate consonants [k] and [p] embedded in meaningless syllables. On each trial, two syllables [ka] and [pa], uttered by the same speaker, were randomly chosen. They were presented in random order, and the participant indicated whether [pa] was in the first or second interval.

In a preliminary experiment, initial threshold estimates were determined for auditory and visual stimuli using an adaptive staircase procedure with a 3 correct-down/1 wrong-up rule, which gives 79 % correct threshold, derived from the average of six turning points. The adaptive method was used to obtain threshold estimates quickly, and more accurate measurement was done in the main experiment.

In the main experiment, discrimination thresholds were measured for five conditions: auditory, visual and three audiovisual conditions, using the method of constant stimuli. In the audiovisual conditions, the ratio of the auditory and visual intensities was kept constant. The three ratios were: A=1.740V, A=V and A=0.575V (forming angles of 30, 45 and 60 degrees in the summation square). They contained more A than V, equal ratio of A and V, and less A than V, respectively.

For each condition, there were six levels of stimulus intensity: 0.1, 0.4, 0.6, 0.8, 1 and 1.4 times the initial threshold, derived from the preliminary experiment. Each level was repeated 30 times. All conditions were randomly interleaved in the experiment.

2.5. Data analysis and fitting

2.5.1. *Psychometric functions*

For each participant, the proportion of correct responses was calculated for each stimulus intensity level, for each condition. The resulting psychometric functions were fitted with a cumulative normal distribution using a Maximum Likelihood criterion with Palamedes toolbox [12]. The threshold value was defined as 79 % correct responses. Then, the auditory and visual thresholds in audiovisual conditions were normalized by dividing them by unisensory auditory and visual thresholds.

2.5.2. *Summation model*

The summation model (Equation 1) was fitted to the normalized audiovisual thresholds by minimizing the squared error between the data and the model. That is, the normalized audiovisual thresholds were expressed as the intensity/contrast of the AV stimulus relative to the unisensory (A or V) threshold. The confidence intervals for *k* values were estimated using a non-parametric bootstrap procedure [13]. In brief, 10,000 psychometric function replicas were generated for each condition and participant by randomly re-sampling (with replacement) the empirical data. Minkowski metric was then fitted to these data, yielding a distribution of *k* values that was used to compute the percentiles of the confidence interval for *k*.

In addition to the individual fits, we estimated the average *k* in this population. For the average estimate, the average of normalized AV thresholds was computed across 16 random samples with replacement from the data containing every participant's thresholds. The best-fitting *k* value was fitted to each dataset. This was repeated 10,000 times to provide a distribution of the *k* values.

3. Results

Examples of psychometric functions from one participant are shown in Fig. 2.

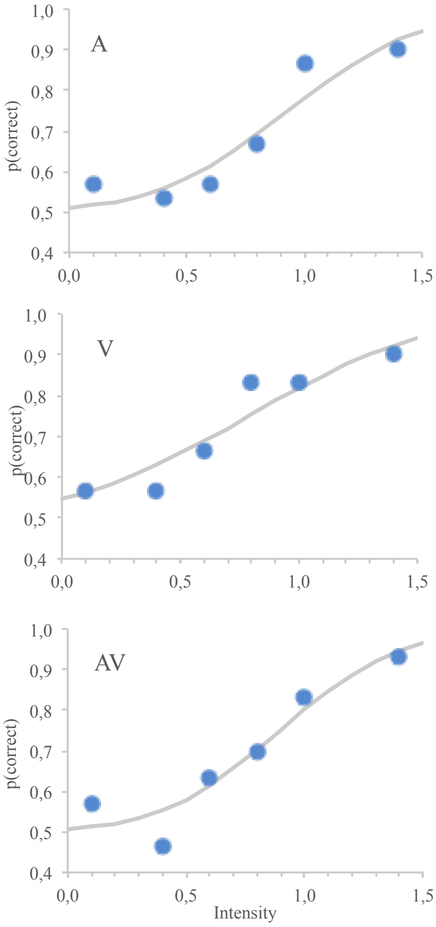

Figure 2: *Psychometric functions: Examples from one participant.*

The summation model was then fitted to the thresholds. Across all participants, the *k* value was 1.7 (Fig. 3). The fit appears rather good.

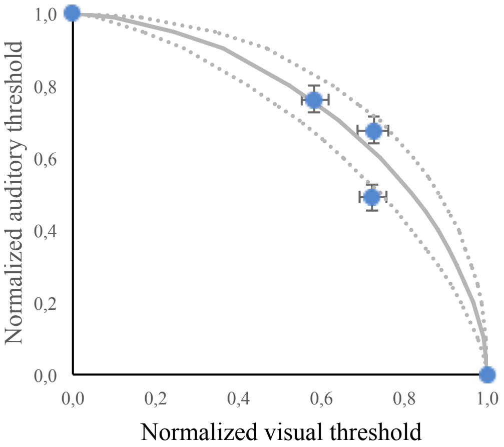

Figure 3: *Summation square for data pooled across all participants. Circles show the normalized audiovisual thresholds averaged across participants. Error bars depict standard errors. Solid line shows the model fit (k=1.7). Dotted line depicts 95% confidence intervals.*

However, there were individual differences in the *k* values (Fig. 4). Summation was closest to linear in 5 participants (*k* ranged between 0.9-1.4), and quadratic in 7 participants (*k* ranged between 1.5-2.4). Weaker summation was present in 4 participants (*k* ranged between 2.6-4.6). The individual 95% confidence intervals did not overlap entirely, suggesting that the individual differences were genuine.

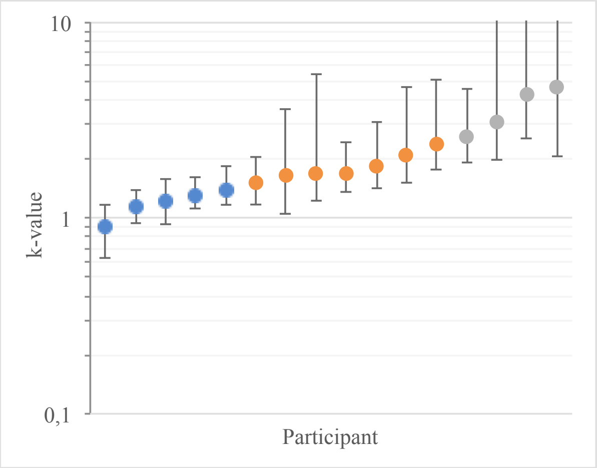

Figure 4: *Individual k values. Error bars show 68 % confidence interval (±SD).*

4. Discussion

Here the established summation model [3-8] was used to estimate the summation of sensory information at threshold in audiovisual speech perception. Auditory, visual and audiovisual consonant discrimination thresholds were measured. Audiovisual summation was estimated by finding the best fit of the Minkowski metric (Equation 1) to the measured, normalized thresholds. The findings were in agreement with previous multisensory integration studies. Summation was quite strong, as would be expected for audiovisual speech perception [8] and for temporally and semantically congruent multisensory perception in general [4-7]. The value of the summation parameter *k* was 1.7 for data pooled across participants. The individual *k* values were most commonly near 2, corresponding to quadratic summation.

Arnold et al. reported linear summation in their six participants, and they concluded that this suggests a physiological mechanism fusing the auditory and visual signals early in perceptual processing [8]. This is in line with the current results in that summation was closest to linear in 5 out of our 16 participants. Since Arnold et al. did not report individual *k* values but instead tested the predictions of Minkowski metric at k of 1 or 2 against the audiovisual data point of each participant, assessment of their findings in relation to the current ones cannot be very accurate. However, in general they report stronger summation than we found in the present study. One difference between the studies is that the white noise employed in our study may have a more potent masking effect than the pixel masks in [8]. Since external auditory and visual noise are statistically independent, that may explain why summation was closer to quadratic in our case. Another possibly contributing factor is the context in which the consonants were presented: embedded in a meaningful, coherent sentence in [8], and in the first phoneme a meaningless, short syllable here. Coherent context strengthens binding in audiovisual speech [14], which may contribute to the stronger summation in [8].

Furthermore, we found some individual variation in the amount of summation. In most studies of audiovisual speech perception, the results are averaged across participants, thus ignoring any individual variation. However, it is known that such variation exists. Individual differences have been addressed using the McGurk effect as an index of integration. In the McGurk effect, discrepant visual speech alters auditory perception (e.g. A[ba] presented with V[ga] is heard as “da” or “ga” [15]. The fewer responses according to the auditory component, the stronger the McGurk effect. The strength of the McGurk effect varies between individuals [16-19], implying that the amount of summation varies. Also, summation in spoken and written words shows some variation between individuals, with *k* values of the summation model ranging between 1.1-1.9 [7]. This variation is in agreement with the current findings.

The current results have implications regarding the explanation of how multisensory summation arises. We found, on average, summation that was close to quadratic. However, some individuals exhibited approximately linear summation, suggesting that the noise source limiting performance was after the summation stage, i.e., post-summation noise dominated. It has been suggested that a mix of noise sources produces performance that varies between linear and quadratic models [20]; see also [21]. We propose that the amount of post-summation noise might vary between individuals. In this scheme the amount of noise at different processing stages determines the strength of summation. If the amount of post-summation noise is high (so that the contribution of unisensory noise is negligible), linear summation is observed. If the amount of post-summation noise is low, unisensory noise dominates and quadratic summation arises. Alternatively, it has been proposed that integration occurs at multiple stages [22]. Thus, the amount of integration at different stages, as well as the pre/post noise level, may vary between individuals. However, since a simpler explanation is generally preferrable, our proposal is more parsimonious.

5. Conclusions

Audiovisual enhancement in speech percption was assessed using the summation model. Audiovisual discrimination thresholds were best fitted with low values of the summation parameter of the model. This indicates strong summation. There was some individual variation in the amount of summation. This suggests that summation is modulated by different sources of noise.

6. Acknowledgements

This research was supported by a travel grant to Kaisa Tiippana from Medicum, University of Helsinki.

7. References

[1] W. H. Sumby and I. Pollack, “Visual contribution to speech intelligibility in noise”, *The Journal of the Acoustical Society of America*, 26, 212-215, 1954.

[2] W. J. Ma, X. Zhou, L. A. Ross, J. J. Foxe and L. C. Parra, “Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space”, *PLoS One*, 4, 3, e4638, 2009.

[3] N. Graham, *Visual pattern analyzers*. Oxford: Oxford University Press. 1989.

[4] G. F. Meyer, S. M. Wuerger, F. Röhrbein and C. Zetzsche, “Low-level integration of auditory and visual motion signals requires spatial co-localisation”, *Experimental Brain Research,* 166, 3-4, 538-547, 2005.

[5] R. Arrighi, F. Marini and D. Burr, “Meaningful auditory information enhances perception of visual biological motion”, *Journal of Vision*, 9, 4, 2009.

[6] D. Alais and D. Burr, “No direction-specific bimodal facilitation for audiovisual motion detection”, *Cognitive Brain Reserach,* 19, 2, 185-194, 2004.

[7] M. Dubois, D. Poeppel and D. G. Pelli, ”Seeing and hearing a word: combining eye and ear is more efficient than combining the parts of a word”, *PLoS One,* 8, 5, e64803, 2013.

[8] D. H. Arnold, M.Tear, R. Schindel and W. Roseboom, “Audio-visual speech cue combination”, *PLoS One,* 5, 4, e10217, 2010.

[9] D. W. Massaro, “*Perceiving talking faces*”, MIT Press, Cambridge, Massachusetts, 1998.

[10] D. G. Pelli, “The VideoToolbox software for visual psychophysics: Transforming numbers into movies”, *Spatial Vision*, 10(4), 437–442, 1997.

[11] M. Kleiner, D. H. Brainard D. G. Pelli, “What’s new in Psychtoolbox-3?”, *Perception*, 36 (ECVP Abstract Supplement), 2007.

[12] N. Prins and F. A. A. Kingdom, “Palamedes: Matlab routines for analyzing psychophysical data”, www.palamedestoolbox.org, 2009.

[13] B. Efron and R. J. Tibshirani, “*An introduction to Bootstrap*”, Chapman & Hall, New York, 1993.

[14] O. Nahorna, F. Berthommier, and J. L. Schwartz. “Binding and unbinding the auditory and visual streams in the McGurk effect”, *Journal of the Acoustical Society of America, 132*(2), 1061-1077, 2012.

[15] H. McGurk and J. MacDonald, “Hearing lips and seeing voices”, *Nature*, 264, 746-748, 1976.

[16] J. L. Schwartz, “A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent”, *The Journal of the Acoustical Society of America*, 127, 3, 1584-1594, 2010.

[17] R. A. Stevenson, R. K. Zemtsov, and M. T. Wallace. “Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions”, *Journal of Experimental Psychology: Human Perception and Performance*, 38(6), 1517-1529, 2012.

[18] J. Strand, A. Cooperman, J. Rowe, and A. Simenstad. “Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity”, *Journal of Speech, Language, and Hearing Research*, 57(6), 2322-2331, 2014.

[19] B. D. Mallick, J. F. Magnotti, and M. S. Beauchamp. “Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type”, *Psychonomic Bulletin & Review*, 22, 1299-1307, 2015.

[20] P. C. Stacey, P. T. Kitterick, S. D. Morris and C. J. Sumner, "The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure", *Hearing Research*, 336, 17-28, 2016.

[21] C. Micheyl and A. J. Oxenham, "Comparing models of the combined-stimulation advantage for speech recognition", *The Journal of the Acoustical Society of America*, 131, 5, 3970-3980, 2012.

[22] J.-L. Schwartz, F. Berthommier and C. Savariaux, “Seeing to hear better: evidence for early audio-visual interactions in speech identification”, *Cognition*, 93, B69-B78, 2004.